MEASURING FACE ICONICITY

by
Tao Zhang

A thesis submitted to Johns Hopkins University in conformity
with the requirements for the degree of Master of Science in Engineering

Baltimore, Maryland
April, 2021

Abstract

In recent years, a significant amount of work has been made to improve the performance of face recognition and verification systems. Although the accuracy of face recognition or verification is very good, the the is room to improve the overall reliability and trustworthiness of the system. The goal of this project is to train a model that can evaluate the iconicity score of any input facial image, and the iconicity score needs to reflect how well a given face represents the identity of the individual in the forthcoming face recognition and verification task. Therefore, we propose a Siamese Multi-Layer Perceptron network to train the face iconicity score using image pairs as the training set. Our training set is generated from the image dataset only with identity information that is already available for the face recognition training task, and our indirectly supervised model does not need any iconicity or quality-related labels or any other information beyond identity during the training process. Our method can help overcome the limitation of the high cost of datasets with quality-related labels.

We verify the effectiveness of our iconicity score by (i) checking the iconic and non-iconic images scored by our model to see if the quality description is consistent with the human option, (ii) exploring the relationship between the iconicity score with other visible factors which will influence the recognizability of facial images, (iii) using iconicity score of faces as a filter to help improve the performance of face verification system by dropping the facial images that do not contain sufficient information to be recognized.

Dedicated to one or more of the various

people or pets or influences in my life.

Touching final thought.

Acknowledgements

For the completion of this thesis, I need to sincerely thank many people for their help to me during my graduate period.

My deepest gratitude goes first and foremost to my advisor Professor Carlos Castillo, for his constant encouragement and guidance. He has helped me a lot in all the stages of finishing this essay. He always gave me inspiration and useful advice. This essay can never reach its present form without his consistent instruction.

Second, I would like to express my heartfelt gratitude to all my friends especially my "Baltimore family", it was really a special year for everyone, we cannot get together in class or at school, and we can't take part in many interesting activities. But my friends give my a lot support, we really enjoy a lot wonderful time together which help me relax a lot. All of you light up my life.

Last my thanks would go to my boyfriend Junyu Shen, who is now in China, we have been separated into different country for near one and a half years, we can't visit each other because of the special situation. But I always believe that the mountains and the seas can be leveled by love. Thank you for standing by me when I was under pressure, see you soon.

Contents

List of Figures

Chapter 1

Introduction

1.1 Problem statement

When people try to recognize someone from a facial image, the "quality" of the facial image will influence not only the recognition accuracy but also the confidence for making the decision. For example, figure 1-1 includes two facial images of the former US president Barack Obama, and the first one is a clear, frontal face, while the second one is a blurred face with a low angle. In the recognition task, it is significantly easier for people to recognize President Obama from the former one, so that it is an iconic facial image which enjoys high "quality".

The quality of the facial image can be estimated from different kinds of standards. The standard of facial images used in the process of government ID image acquisition such as ISO/IEC 2382-37 [1] takes several factors into consideration such as face/eye location, blurring, illumination, if looking away. However, such a standard might be too strict for most of the verification tasks whose images are captured from unconstrained scenarios. In the last few years, face recognition and verification systems have advanced rapidly. The architectures of the deep convolutional neural networks (DCNNs) update from AlexNet [2], VGGNet [3], to ResNet [4], and in the meanwhile, several large-scale datasets such as VGGFace [5], UMDFaces [6], IJB-C dataset[7] have appeared and enable training and evaluating large neural networks. The state-of-art face recognition systems can extract more effective features from face images which are more stable between different poses, yaw, illumination, etc. To estimate the quality of facial images used in face recognition tasks, we believe the estimation should be based on the performance of the face recognition system and the result quality score need to reflect the machine-recognition potential of the facial image. Thus, it would be very helpful to have a method, which requires minimal supervision, that can evaluate the quality of images based on the traditional face data set that only contains the labels about identifies, and does not have any other quality-related ground truth. So we propose iconicity to describe such kind of quality evaluation.

(a) President Obama (b) President Obama

Figure 1-1. (a) high quality and (b) low quality images of president Obama

1.2 Face verification

Facial image quality evaluation can exert a positive effect in the face verification task. Face verification is the task of comparing a candidate facial image to the reference face, and then determine if they are from the same individual. In this project, we first detect and crop the face area from the complete image with face detection and face alignment, and then use the pre-trained DCNN, which trained based on the crystal loss function, to extract feature vectors from each face based on [8]. This process will be explained more later. Besides, on the crystal loss optimized descriptor space of face representations, the cosine similarity of facial image feature pairs can provide an estimation of the angular separation between these facial images, so that we use the cosine similarity between feature vectors of facial image couple as the verification score of each image pair, which can be shown as figure 1-2. Based on the verification score, the decision about whether the two images are from the same person or not will be made.

Figure 1-2. Verification scores for matched pairs

The verification score can measure the likelihood that two images are the same individual. However, the score might be imprecise due to many confounding factors, and one of the main confounding factors is the quality of the images being compared and the quality difference between reference and test images. In general, the reference facial image is obtained in a constrained situation so that it enjoys high quality, but the test data will be more varied. As a result, the estimation of facial image quality can help filter the test facial image or help compensate for the influence of the image quality to improve the performance of the face verification task.

1.3 Face quality assessment

1.3.1 Application Scenario

According to face image acquisition scenarios, the facial images can be divided into two scenarios: the facial images obtained from constrained and unconstrained scenarios. Under the constrained scenario, the subject of facial images will be cooperative and look at the camera, the illumination, the background will be set to ideal conditions in advance. Therefore the quality assessment, in this case, can be used to ensure that all the requirements are satisfied during image acquisition. Under the unconstrained scenarios such as images captured from station monitoring, the images are obtained without any cooperation from the individuals or any restrictions for the environment. The quality of facial images here is used as a filter to reject the facial images whose quality is too low to be recognized by machines, and in this way, the quality scores can help improve the performance of face verification systems.

1.3.2 Acquisition method

There are many methods to obtain the quality scores/level of the facial image, probably the most straight forward one is the method that trained based on the data set which require the ground truth of the quality such as [9–13]. These methods are viable if the quality-related labels for large image datasets are available. However, the approach we propose does not need that kind of quality-related information. We infer the iconicity scores from the images themselves based on the data set that only contains the identity label, and the optimization of the model is running based on the identity label and the performance of the data set in face verification systems. Our project proposes a quality assessment algorithm that does not need any extra information about the image quality because the subjectivity of image evaluation will influence the reliability of the quality ground truth and obtaining a data set with the quality label is difficult and costly.

1.4 Iconicity of facial images

The iconicity score of the facial image is used to describe how well a given facial image represents the identity of a given person in the forthcoming face recognition and verification task [14]. A face image with a high iconicity score is easy to be recognized and matched with the template image. If we have the iconicity labels for a set of faces, we could use the machine learning method to train a system that generalizes to unseen faces. However, the quality of facial images and the confidence of recognition is subjective for each estimator, which makes the judgment of quality unreliable, and it is difficult and costly to obtain the ground truth of the iconicity score provided by people. So obtaining iconicity scores with quality-related labels supervised training method is not the direction we will pursue in this work. Besides, evaluating face iconicity with the unsupervised method by assigning iconicity label (iconic/non-iconic) based on if the image is correctly classified during clustering task is not the best way either, because the model trained with clustering cannot be generally used to predict the iconicity score of the identities who are not included in the training data. Hence, the iconicity of a facial image can only be obtained based on the inherent feature of the facial image itself, which means that the model to compute iconicity score can only be trained based on the face images themselves without any other quality-related information.

Under such a definition of iconicity, the output iconicity score will be a quantitative index to represent how much information about identity verification the facial image contained, and such an index can help predict the face recognition performance without truly performing the face recognition task itself.

1.5 Predictive confidence for image pairs

During the face verification task, we compute the similarity between two facial image features to determine whether two facial images are from the same identity, and such a rule is based on the assumption that both facial features contain effective and sufficient information for recognition. However, such an assumption is not always satisfied during the test process in real-life scenarios because the input images might be blurred, profile, or even not a face at all. As a result, we need an indicator to represent the reliability of the verification result or even use such an indicator to compensate for the influence caused by insufficient information (low iconicity) on their similarity. The indicator can be defined as predictive confidence for image pairs.

From the human point of view, Natu et al. has proposed that Recognition confidence of novel facial images of the known individuals improved with increasing familiarity with faces in [15], and he also found in [16] that the familiarity can increase people's ability to recognize faces in bad viewing conditions. One can always recognize a familiar friend with high confidence quickly even from a low-quality facial image, but it might be more difficult to match an unfamiliar one. One of the reasons for that is the template image for a familiar friend in one‘s memory enjoys high quality because they have met many times and the memory has been reinforced over and over again. Such a phenomenon also happens to machine recognition. The template image which enjoys high iconicity will enjoy higher predictive confidence while matching the test image than the low iconicity template, and the predictive confidence depends on both the template image and the test image. Therefore, in our project, we use the multiple of the iconicity scores of single facial images in an image pair to represent the predictive confidence of the image pair.

1.6 Applications

The measurement of iconicity can be applied in various areas and situations. Among those applications, the most basic one is to let a machine help choose the best facial image to describe an individual. After obtaining a set of facial images for an individual, we can input those facial images into our model and pick the one with the highest iconicity score. Furthermore, there are more other applications of the iconicity score.

1.6.1 Acquisition process feedback

The iconicity score can be used in the process of facial image acquisition. During the process of image acquisition, all of the images with iconicity scores lower than a threshold will be rejected in real-time, and the individual will be asked to redo the acquisition process to satisfy the requirement. Besides, it can also be used for giving feedback in other kinds of image acquisition processes such as the process of uploading, transmitting and printing facial images.

1.6.2 Weighted template fusion

In a face-related task, the iconicity score can be used as the weight for the fusion of different decisions such as the emotion detection, gender classification, obtained from the same model with different input samples [17, 18], which means that more iconic facial images have higher weights when building the template than the non-iconic ones. We assign weights to images in this way because the images with more information can always help the system make more accurate decisions.

1.6.3 Evaluation of face verification systems

Most of the face recognition and verification system can work well on the data set consisting of iconic facial images. To compare their capability of extract features that are invariant with different conditions like different yaw, expression, etc. Those systems should be test in real life (unconstrained) scenarios, and the difficulty of the verification and recognition task depends on the "difficulty" of the test data set which can be evaluated with the iconicity score. Evaluation of the face verification system with taking the difficulty of test set into consideration will be more trustworthy.

1.6.4 Improving the performance of face verification systems

The facial image iconicity can also be used as a filter to reject the images, which are the frames captured from videos, with quality and iconicity lower than the previously set threshold, and the accuracy of the recognition and verification system tests on the remained image data will be increased. In this application, the iconicity score can be used to improve the performance of the facial image verification system by keeping only the effective facial image/ effective facial image pairs.

1.7 Related work

Measuring the iconicity of facial images is closely related to the face image quality estimation. The proposed literature on image quality estimation can be generally divided into two categories: one is the technique without deep learning models, the other one is the technique based on deep learning models.

1.7.1 Non-deep learning techniques

There are many techniques estimate the quality of facial images with the non-deep learning methods, and they are based on several visible factors such as pose, blur, positions of eyes, etc. For example, there are techniques specifically considered the image requirements given by the standards ISO/IEC 19794-5 [19], ICAO 9303 [20, 21], and ISO/IEC TR 29794-5 [22], etc. These standards are under development as an international standard. Facial images for government-issued ID documents are usually acquired under controlled conditions. Those facial images are under strict quality requirements for several understandable factors such as recommended ranges for the pitch/ yaw /roll of the head. Additionally, the authors of [23] evaluated image quality based on contrast, brightness, focus, illumination, and sharpness, and Anish et al. [24] proposed the BRISQUE algorithm based on the assumption that the high-quality facial images should be a 'natural' image, and the naturalness is defined as the deviation of the normalized luminance distribution from its Gaussian estimation.

All of these methods are using several specific features of facial images to give the facial image the quality assessment, therefore, none of them can be applied to larger image data set with more variable conditions. In order to generalize the quality score obtained from the trained model to a larger and more variable data set, people start to use the deep learning-based method to train models for generating the quality scores of facial images.

1.7.2 Deep learning based techniques

The image quality assessment approaches based on deep learning includes two different kinds of model: supervised model and indirectly supervised model.

1.7.2.1 Supervised model

Techniques based on the supervised model require manually extracted quality-specific information to be the ground truth during the training process. For example, Zhang et al. [10] has created FIIQD, the "Face Image Illumination Quality Database", with subjective illumination quality scores for 224, 733 images with 200 different illumination patterns. Based on this dataset, the ResNet-50 model was trained to evaluate the facial image quality (illumination quality). Besides, Yang et al.[11] presented "DFQA", an FQA CNN based on SqueezeNet. For training, 3000 images were first manually annotated with ground truth quality scores using a defined set of rules to increase the quality score's objectivity/subject independence. Then those labeled images were used to do transfer learning based on the pre-trained SqueezeNet to predict the quality scores of input images.

According to those experiments, the CNN/DNN based supervised quality assessment models require supervision information about the quality scores. However, obtaining the data set with the ground truth of the quality information is expensive and time-consuming. Moreover, such a kind of data set is always with limited size and can only consider several factors. As a result, several approaches which try to train the indirectly supervised model come up, and they do not rely on the quality information of the training data.

1.7.2.2 Indirectly supervised model

In order to get rid of the limitation from the lack of the data sets with quality annotated, several techniques use only the inherent features of input images and an identity labeled data set to obtain the quality scores and iconicity scores. For example, one of the interesting work is [25], this paper use the norm of the facial image feature vector to measure the quality of the input face, and it demonstrated that the low quality images always accompanied with lower norm value of the feature vector, but such an approach cannot be used in the network which will normalize the feature vector. Besides, another approach to get the quality score is by learning the set representation, which was proposed in [26] . The idea is to compute the set representation by calculating the weighted average vector among all of feature vectors of one individual, and the weights are the variable which will be obtained from the deep network. As a result, during the process of optimizing the classification, the set representation will try to increase the weight of the high-quality face and decrease the weight of the low-quality one, so the final weight can be treated as the quality estimation. However, such a result at times suffers from over-confident predictions, which means that most of the images for one individual will get high quality score.

Moreover, there are also several methods that train networks based on the cosine similarity of image pairs, which means that such methods train the iconicity score of images based on their performance in face verification tasks. For example, Dhar et al. proposed a method to estimate iconicity in [14] with assuming that if a facial image cannot be matched successfully with images of the same identity, it is a non-iconic image. Based on this assumption, the paper [14] maps the hypersphere of images for verification to another one where the length of the feature vectors are the iconicity of this image feature, which is the output of the network. Then it used the binary label of image pairs (1 for genuine pair, 0 for imposter pair) to train the objective function and enable the network to estimate the verifiability of a feature. However, in this network, the soft margin of verification score is difficult to decide since it needs to be changed according to different data set and the result iconicity scores for images mostly concentrate around 1, which means that it suffers from over-confident results. Another method based on face verification performance is proposed by Xie et al.in [27], this paper proposed the Predictive Confidence Network (PCNet) which assumes that the verification scores of image pairs are equal to the minimum of the two image' confidences in the pair. Therefore the objective function of PCNet is trained to push the minimum confidence to verification score, and the output scalar may indicate the likelihood of the face being identifiable by a verification system. The only problem with this method is that the basic assumption of "loser takes all" can not explain the verification confidence well, especially when we apply the iconicity score to unbalanced facial image pairs (which means most of the facial image pairs consist of one iconic and one non-iconic facial image). That's because this assumption will ignore the influence of the more iconic facial image on the final image pair verification score by equaling the facial image pair verification score to the minimum iconicity score of the facial image pair.

In our project, we propose an approach to train a Siamese MLP network to predict the iconicity of facial image for any input facial image based on the methods from [14, 27], the result can be roughly shown as figure 1-3. The image sets we need are the data sets that contain facial images with varied acquisition conditions, varied poses, and different noise levels, and our method only requires identity labels that are already available for face recognition training tasks. With this kind of image set, we can generate the training set that consists of facial image pairs from the same individual (matched image pairs) to train our network. Our method does not need any extra information about the quality or iconicity ground truth of the training images, and our method does not depend on the norm of the feature vectors of facial images which cannot be used when the feature extractor normalizes the feature vectors. Therefore the main contribution of our method is that we overcome the limitations of the supervised model of the high cost of data set with quality ground truth and our method can also be used in the network which will normalize the feature vectors of facial images. Besides, our approach for evaluating the iconicity of facial images is more appropriate to be used in the situation when most of the test images in the face verification task are non-iconic ones, which means that most of the face verification image pairs are unbalanced.

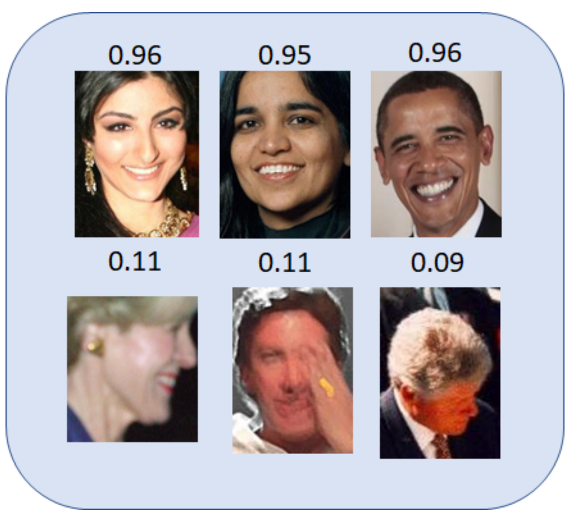


Figure 1-3. Images with different iconicity scores

Our method to predict the iconicity of facial images is described in chapter 2. In chapter 3, we describe different metrics to evaluate our network and result and we also propose several applications of our iconicity score based on these metrics.

Chapter 2

Approach

In this chapter, we propose the Siamese multi-layer perceptron (MLP) network, which receives as input the feature vector of the facial image and produces as output a scalar representing if the input image contains sufficient information to be recognized by the face verification system. Here we use deep feature representations of faces instead of the raw image to be the input of the Siamese MLP network. In order to extract the 512-d feature vector, we first apply face detection algorithm to the raw image and crop the facial area from the image, then we use a pretrained DCNN to extract the 512-d feature from the face area of the raw image. There are two steps during the training of the network: first one is to generate facial image pairs from same individual and then calculate the verification score of the matched image pair in a simple and scalable way. Here we use the cosine similarity to be the verification score. Second one is to disentangle the pairwise verification score and optimize the objective function by updating the parameters of each sub-network until converge to obtain scores of single faces. In such a way that those inherent features contained in the images will help update and optimize the network, and the network will be able to predict the iconicity of the single input image even if the input individual isn't included in the training data set.

2.1 Computing deep descriptors of face images

In our network, we use deep feature representations of faces to be the input, which are extracted from the architecture proposed in [8]. This paper describe a deep learning pipeline for unconstrained face identification and verification, and achieves state-of-the-art performance on several benchmark datasets. The face identification and verification system proposed in this paper includes a novel face detector called deep pyramid single shot face detector (DPSSD), which can detect faces with large scale variations (especially tiny faces) fast. The DPSSD starts with the Single Shot Detector (SSD) which is trained on the truncated VGG-16 [3] network for object detection. Based on the SSD which will generate a fixed number of bounding boxes and corresponding scores for the presence of faces, this detector additionally adds convolutional layers to detect objects at multiple scales at the end, so that such a modified SSD is able to detect tiny faces efficiently.

Besides, the loss function in this paper for the tasks of face verification and identification is a new loss function called crystal loss, which can restrict the feature descriptors to lie on a hypersphere of a fixed radius, thus minimizing the angular distance between positive subject pairs and maximizing the angular distance between negative subject pairs. The crystal loss function can be shown as:

$$L_s = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{e^{{W_{y_i}}^T f(x_i)+b_{y_i}}}{\sum_{j=1}^{C} e^{{W_j}^T f(x_i)+b_j}} \tag{2.1}$$

where M is the training batch size, x_i is the ith input face image in the batch, $f(x_i)$ is the corresponding output of the penultimate layer of the DCNN, y_i is the corresponding class label, and W and b are the weights and bias for the last layer of the network which acts as a classifier.

2.2 Assumptions

Our method is proposed based on the basic ideas of PCNet, we also design our objective function based on the verification scores of the matched image pairs. However, the biggest difference between the objective function of our network and PCNet is due to our different assumptions.

Different from the assumption of "loser takes all" in PCNet, our method is based on the assumption that the verification score of the matched image pair is decided by both of images in the pair, which means that if the final verification score is less than 1.0, then it might because both of the image miss some information for recognition. We raise this assumption based on how human match two images. For example, if there are two pairs of images, one of them consists of one iconic (high iconicity score) image and one non-iconic (low iconicity score) image, and the other one consists of two non-iconic images. People will match the former with more confidence than the latter. We can see an example in figure 2-1

(a) Image pair that includes iconic and non-iconic images

(b) Image pair that includes only non-iconic images

Figure 2-1. Comparison between different image pairs

In order to represent the relationship between the iconicity score of single image and the predictive confidence of image pair, we assume that the predictive confidence of image pair is the multiple of the iconicity scores of each single image in this pair, so that both images' iconicity/quality will make influence to the final confidence scores.

2.3 Network structure

The structure of our network is based on the Siamese neural network, which is inspired by [14] which also takes use of the Siamese neural network to train the iconicity score, but different from the hinge loss function it uses which tries to increase or decrease the output score based on if the image can be matched correctly with other images of the same person, our model have more meaningful training objective that our iconicity score is trained to satisfy that the multiple of the iconicity scores of images will be equal to the verification score of the image pair. Moreover, because of this objective, our model are trained only with the genuine image pairs while the model in [14] was trained with both genuine and imposter image pairs.

Siamese neural network consists of two sub-networks who share the same weights and same architecture. These two sub-networks work in tandem with two different input, which, in our model, is two feature vectors from a matched pair. The comparable output of the two sub-networks will be used as variables to compute the objective function. During the training process, the objective function will be optimized by updating the weights of the sub network until coverage, and during the test process, we will use just half of the Siamese network to get our target variable: iconicity score of the single image. The structure of Siamese MLP network we built here is showed in figure 2-2

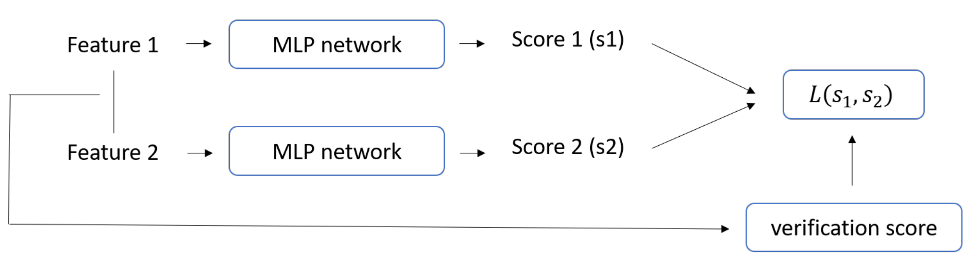


Figure 2-2. Structure of the Siamese MLP network

The input of the MLP network is a 512-dim feature vector, and there are 4 hidden layers into the MLP network, which consists 512, 256,128,64 hidden unites respectively.

The first three layers are followed by the non-linear activation SeLU, and each layer is connected by full connection. The nodes of last layer are summed up and activated to calculate the output result, which is scaled between 0 and 1 with sigmoid unit. The output scalar is the iconicity score of the input feature. The structure of the MLP network is showed in figure 2-3

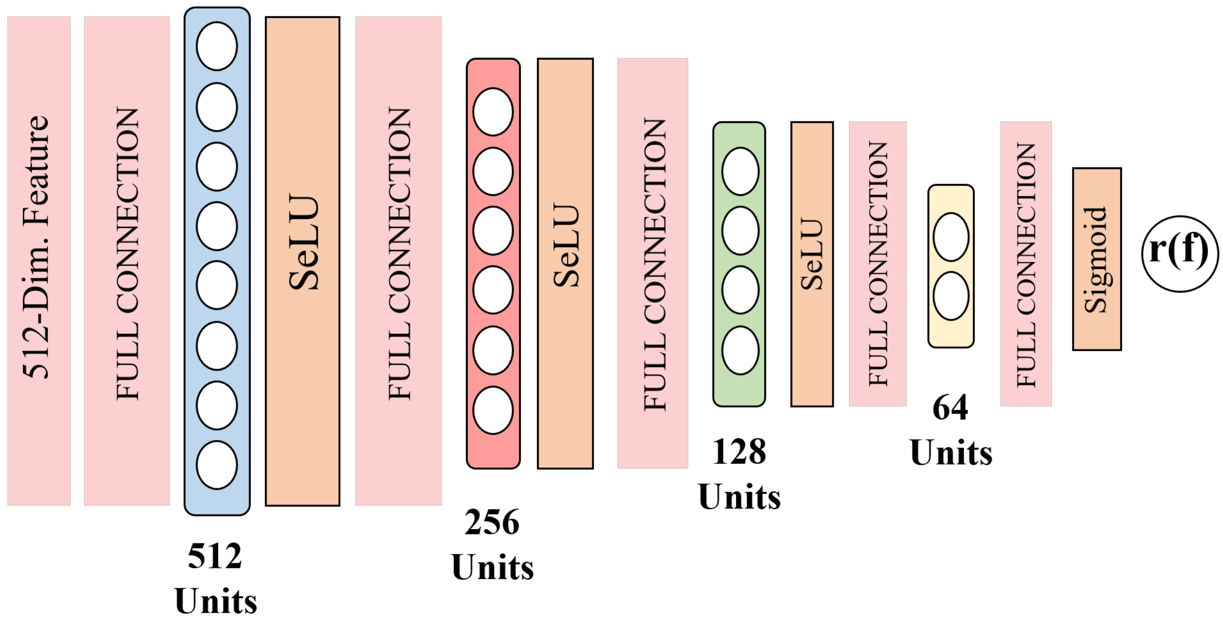


Figure 2-3. Structure of the Siamese MLP network

Moreover, during the test process, only half of the Siamese MLP network is used, which means that during the test process, we can input one feature descriptor of a facial image, and get the iconicity score of that single image as output.

2.4 Objective function

Based on the assumption and structure description above, we can know that our model is trained to output the iconicity score for single image. After building the Siamese MLP network, we need to use the output of the two sub-networks, which are the iconicity scores of two input facial images to generate the objective function, so that we can make the iconicity score represent the capability of the image to be recognized by optimizing the objective function. While processing face verification task, human always has high confidence to claim that a pair of face are from the same individual when these two facial images are of high similarity score. So that we can assume that in the ideal condition, the predictive confidence, which is the multiple of images' iconicity scores from the matched pair, should be equal to the verification score of the matched pair, and if we can induce the predictive confidence score to be equal to the verification score, the meaning of iconicity score can be the effectiveness of the facial image in face verification task. As a result, the objective function of the network can be built as following:

$$L(s_1, s_2) = |s_1 s_2 - y|^2 \tag{2.2}$$

s_1, s_2 is the iconicity score of two input image feature vectors. y is the verification score of the image pair. During the process of training, the objective function will be minimized so that the difference between verification score and predictive confidence will be minimized.

During the training process, the Siamese MLP network will take in a pair of feature vectors from a matched image pair (from the same individual) and output the iconicity score $s(f_1), s(f_2)$ respectively. The s(f) will be initialized as a random scalar between 0 to 1, and as the training process running, the process of optimization will update the parameters (weights) of the sub-network, and the s(f) will be optimized in a way that push the multiple of iconicity scores for each pair to the verification score, which will help the iconicity capture the identifiability of features with respect to the entire data set.

However, because the objective function of the Siamese MLP network is based on the multiple of the input images' iconicity scores, the process of optimization will update the parameters (weights) in the same way for these two subnetworks, so it will cause the same influence to these two input images. But such a issue will be solved with the correct way to generate training set.

2.4.1 Training set generation

While generating the training set (image pair set) from the image dataset which is a subset of the IJB-C dataset (we call the image set IJB-C-sub), we choose each image of one individual first, and than couple it with all of the rest images from the same individual. So that we can make sure each image has been paired with any other options from the same individual to generate matched image pair. Only in this way, we can make sure that the iconic image will enjoy higher iconicity scores than the non-iconic one.

Based on the way we generate the training set, we can know that for an iconic image, it will be coupled with other iconic images or non-iconic images. When it meets with the iconic images, the verification scores will be high, and the optimization process will tend to increase the iconicity score of the images; When it meets with the non-iconic images, the verification will be relative low, so the iconicity score will be decreased. However, for a non-iconic image, it will also be coupled with both iconic and non-iconic images, but no matter what kind of image it coupled with, the verification score will be relatively low, and the iconicity score of it will be decreased. As a result, the iconic image will enjoy higher iconicity score after the over all training process, so that our objective can be realized.

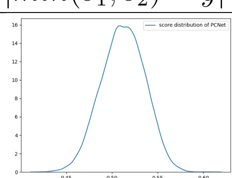
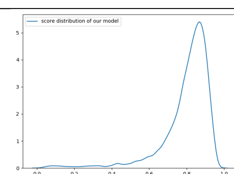
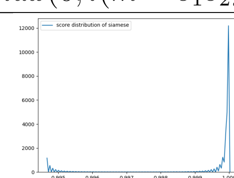
2.5 Implementation

In this project, we built our model Siamese MLP network with Python 3.7.4 and based on PyTorch 1.6.0. Besides, in order to compare the results, we also implemented the PCNet [14] and Siamese network [14] according to the descriptions in each paper with the same version of pPthon and PyTorch. We used the subset of IJB-C Dataset [28] to train the Siamese MLP network. Our dataset is a subset of the IJB-C dataset which includes about 11292 images from 1813 different individuals, these images include both iconic and non-iconic samples. We divided the dataset into the training set and test set in an 8:2 ratio.

2.6 Comparison

The following table 2-I shows the comparison among our Siamese MLP network, PCNet in [27] and Siamese network in [14], for the loss function, s_1, s_2 is the iconicity score of two input image feature vectors. y is the verification score of the image pair. During the process of training, l is the label of if the image pair are genuine pair, m is the margin:

Table 2-I. comparison among different models

content	PCNet[27]	Siamese MLP network(our model)	Siamese network[14]
Assumption	loser takes all	verification score is decided by both images in pair	None
Input	image	feature vectors	feature vectors
Structure	ResNet18 [29]	Siamese MLP network	Siamese MLP network
Training data	genuine image pair	genuine image pair	genuine and imposter image pair
Loss function	$L(s_1, s_2) = \lvert min(s_1, s_2) - y\rvert^2$	$L(s_1, s_2) = \lvert s_1 s_2 - y\rvert^2$	$L(s_1, s_2) = max(0, l(m - s_1 s_2 y))$
score distribution			

Chapter 3

Experimental evaluation

The ideal iconicity estimation should be able to solve various face identity-related tasks and should be consistent with human perception. In this chapter, we carry out several experiments to be the metrics for our iconicity scores. We prove the validity of our iconicity score by (i) plotting the facial images with different iconicity scores to see if the evaluation is consistent with human's opinion, (ii) checking the relation between the iconicity score and other visible factors like yaw, pitch, roll, (iii) verifying the effect of using iconicity score on the performance of face verification system. Moreover, we also compare the effect of our iconicity score and the iconicity score obtained from PCNet on the performance of the face verification system to show the superiority of our iconicity score.

3.1 Datasets

We use the subset of IJB-C Dataset [28] to train the Siamese MLP network. The IJB-C-sub dataset includes about 11292 images from 1813 different individuals. These images include both iconic and non-iconic samples, and a portion of these images are frames from videos captured under the unconstrained scenarios with large variations in viewpoints, illumination, blur and image quality, etc. The IJB-C dataset is generally considered one of the most challenging public benchmarks for face recognition.

3.2 Metrics

An effective iconicity score should be a comprehensive indicator of the test facial image and should be consistent with human judgment. Besides, the most important thing is that the iconicity score should be able to reflect the identifiability for machine and face recognition system, which means that the iconicity needs to be able to show if the test facial image contains enough information to be recognized.

3.2.1 Grading images

The basic application of measuring the iconicity score is to grading facial images. Therefore, we can plot facial images with different iconicity scores to see if the system's judgment is consistent with human perception. Figure3-1 shows 10 iconic facial image and 10 non-iconic facial image:

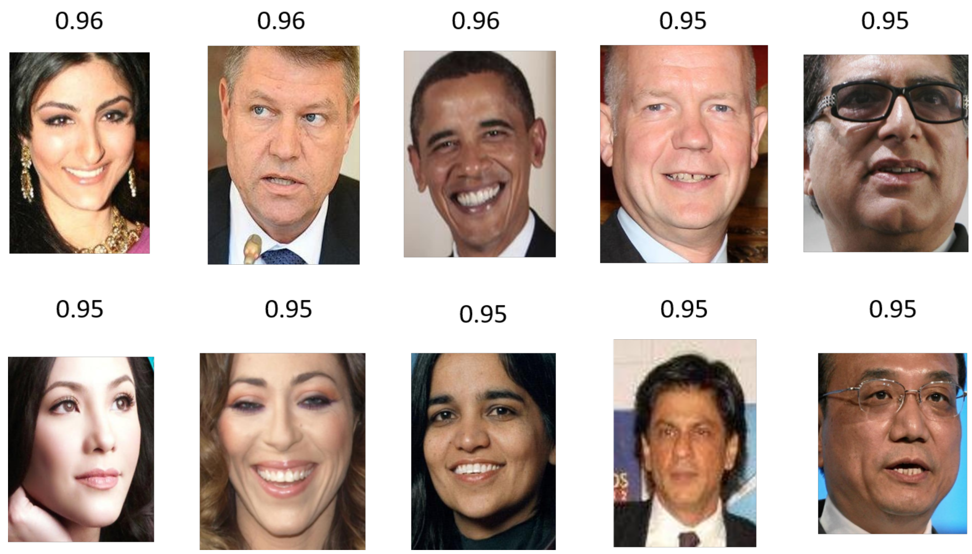
According to figure3-1, we can easily know that the iconicity score can reflect the difficulty for people to recognize facial images. Therefore we can say that our result is consistent with human perception.

3.2.2 Relationship with other factors

To verify the effectiveness of the iconicity score obtained from our model, we explore the relationship between the iconicity score from our model and other factors. These factors include the facial factors that are visible to people and the key points captured by face recognition systems.

3.2.2.1 Relationship with pitch, yaw and roll

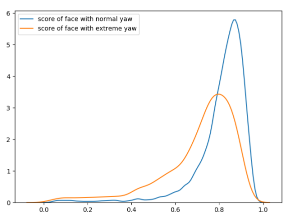
It's easy to know that the most recognizable faces are the frontal faces, and when the pose of faces becoming extreme, the difficulty to recognize increases for both people and machines. Therefore, in order to verify the effectiveness of our iconicity score, we will explore the relationship between the iconicity score and the yaw, roll, pitch



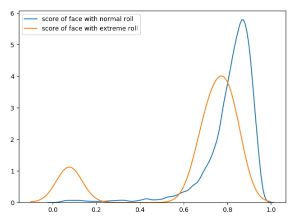
(a) Iconic images

(b) Non-iconic images

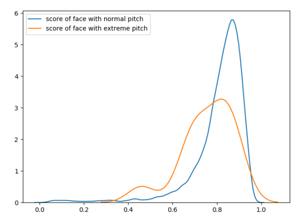
Figure 3-1. Images with high(a) and low(b) iconicity score of faces. We compute the values of these parameters for faces with the all-in-one ConvNet [30]. Then, we plot the distribution of the iconicity scores for different kinds of poses. We define the yaw, roll, and pitch to be normal pose when the angles are between $-45°$ to $45°$, and to be extreme pose when the angles are higher than $45°$ or lower than $-45°$. Besides, when we explore the relationship between iconicity scores and one of these factors, the other two factors are limited to the normal pose. The distributions of iconicity scores for different poses are shown in figure 3-2.



(a) Iconicity score distribution for different yaw



(b) Iconicity score distribution for different roll



(c) Iconicity score distribution for different pitch

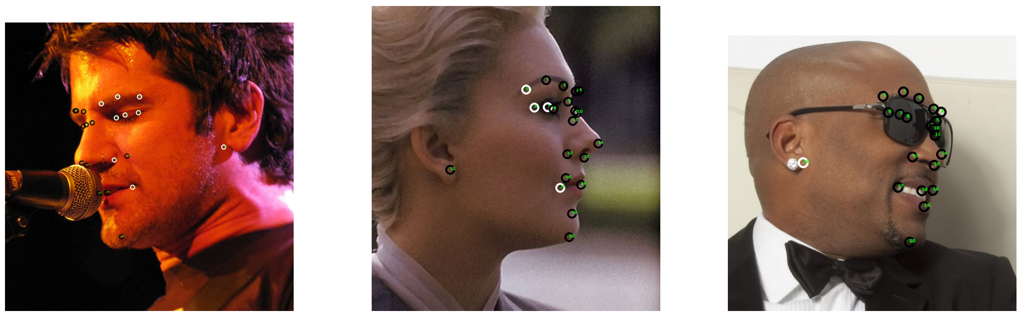
Figure 3-2. Iconicity score distribution for different pose

It can be known from the result that the Siamese MLP iconicity scores correlate well with pose variation, because that the iconicity score captures the pose features of the face and reflect the influence caused by different poses in a way that is in line with people's instincts.

3.2.2.2 Relationship with key points

Facial key points are important features and indicators of facial images. Key points detection involves detecting key points like centers and corners of eyes, nose tip, and these key points can reflect the recognition information contained by a facial image. If we can detect all the key points correctly and completely with high key points visibility from a facial image, we can say that this facial image is an iconic one because it contains sufficient information to be recognized.

Therefore, in order to verify the relationship between the key points and the iconicity score, we plot the key points on the iconic and non-iconic facial image to see the difference. Besides, the key points of high visibility will be a white circle and the low-visibility key points will be a black circle, the results are shown in figure 3-3.



(a) Image with score 0.982 (b) Image with score 0.780 (c) Image with score 0.042

Figure 3-3. Key points of images with different iconicity scores

The result tells us that in the image with a high iconicity score, the key points are always correctly captured and there will be relatively more “white” key points which means more key points are with high visibility. However, in the image with a low iconicity score, the key points always can not be correctly captured. For example, in the image (c) in figure 3-3, the key points which should be captured on eyes can not capture eyes correctly because the eyes are occluded by the sunglasses, and the eye key points here are fake. Besides, almost all the key points on the image (c) are black key points with low visibility. Therefore we can say that the iconicity score from our model correlates well with the ability to capture the key points, and such a fact helps verify the effectiveness of our iconicity score.

3.2.3 Improvement of verification performance for specific use

3.2.3.1 Effectiveness test based on face verification task

The iconicity score is used to describe the identifiability of the image, therefore one of the important metrics is based on the face verification task. According to our assumption, the predictive confidence score (verification score) of the image pair can be calculated by the multiple of two images' iconicity scores. The effectiveness of iconicity can be evaluated with image-to-image verification.

The performance of the face verification system can be reported by the receiver operating characteristics (ROC) curve, which is True Accept Rate (TAR) vs. False Accept Rate (FAR). The ROC curve is built by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. In the face verification task, the threshold variable will be the cosine similarity between two feature vectors. Therefore, we can know the better performance always enjoy a higher true accept rate, especially in the low false accept rate area.

In order to test the iconicity score, after we obtain the iconicity score of the test images, we will couple them randomly to input to the verification system. Before that, we will reject a portion of the image pairs with the lowest predictive confidence score (multiple of the iconicity score of images), and then, we will perform the validation task and calculate the TPR and FPR to build the ROC curve on the remaining images. The ROC curve obtained on 90 percent, 80 percent, and 70 percent of the original test set is shown at figure 3-4:

According to these ROC curves, the true accept rate is higher with the same false positive rate when the rejected image pairs increase, especially in the low false accept rate area. It means the performance of face verification goes better when we reject more bottom image pairs. Such a tendency indicate that our iconicity score can predict images' identifiability for machine effectively and it is directly related to how difficultly the image being matched because the uninformative images are being rejected first here.

Besides, we can also report the true accept rate (at the fixed false positive rate) versus rejection curves for image-to-image verification to prove the effectiveness of the iconicity scores or the predictive confidence score. Figure 3-5 shows the TAR versus the different ratios of rejection.

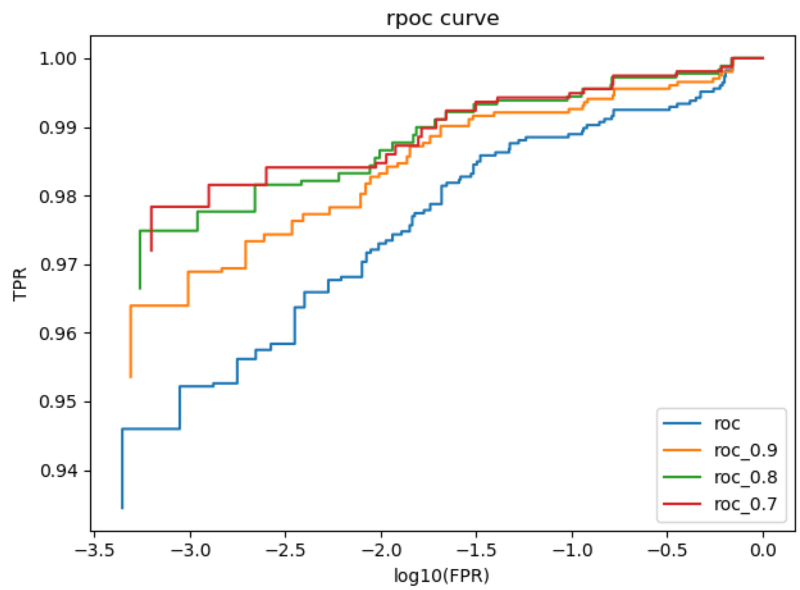


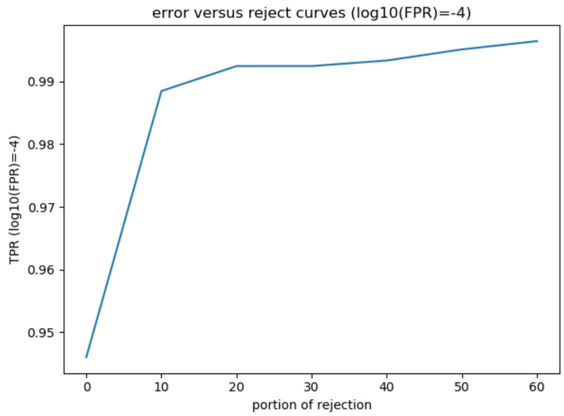
Figure 3-4. ROC curve our model tested on image pairs with top 90,80,70 percent iconicity score

The curves indicate that the iconicity estimation works well because the verification performance increases (the error decreases) monotonically as more images are rejected (as this indicates that the uninformative images are being rejected first).

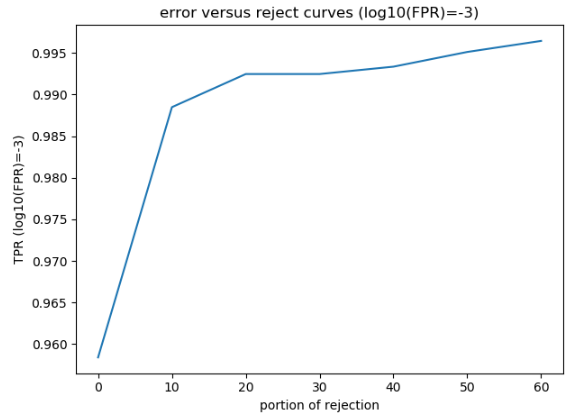
3.2.3.2 Performance comparison in specific applications

Baseline model

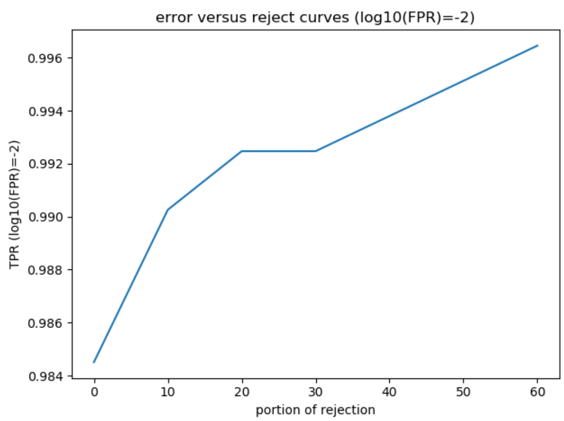
There are several face quality/iconicity estimation algorithms to score the facial images, and one of them called PCNet [27] works really well when applied to the verification task. According to the paper, PCNet can help improve the performance of the verification task by rejecting the non-iconic images, and rejecting the same portion of the dataset based on the score obtained from PCNet will help improve FPR more than the score obtained from other models like Multicolumn Networks (MNet) [31] and Face-QNet (QNet) [32]. PCNet model is built based on the assumption called "loser takes all" which means the verification score of the image pair is decided by the one with the lower iconicity score. Here we will compare the effect of our iconicity



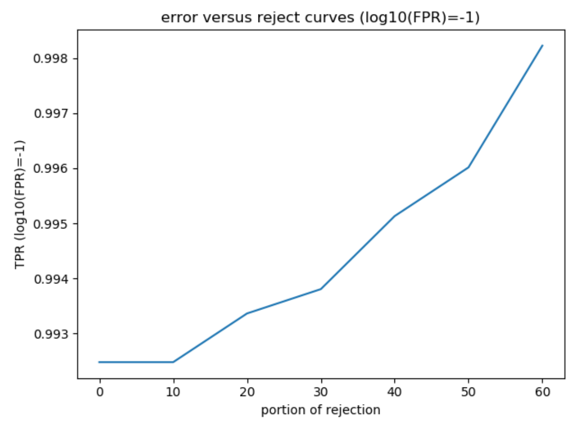
(a) TPR versus different rejection ratio when log10(FPR)=-4)



(b) TPR versus different rejection ratio when log10(FPR)=-3)



(c) TPR versus different rejection ratio when log10(FPR)=-2)



(d) TPR versus different rejection ratio when log10(FPR)=-1)

Figure 3-5. TPR versus different rejection ratio for different FPR score with that from PCNet.

Applications

Face verification is used in a lot of situations in our real life, and a large portion of the applications can not guarantee that the test images are obtained in an ideal scenario. Therefore, we will always see that while processing the verification task, one image of the image pair is clean, captured in constrained condition, but another one might be blurred, in bad condition, or maybe the test one is a profile while the reference is the frontal face. For example, this always happens when we try to apply the iconicity score to catching prisoners. The image of the prisoner on the order for arrest is always in good condition, most of them are obtained from the government id information, but the one we use to test may be captured from the monitor at the railway station or other public places. The frames obtained from the monitor videos are always blurred and the expression, yaw, illumination is not constrained at all, besides, the prisoner may even occlude themselves on purpose. As a result, how to improve the verification performance in this kind of situation, and how to increase the accuracy while the image pairs are unbalanced (one is high-quality and the other one is not) are interesting and important questions.

When most of the facial image pairs in verification tasks are unbalanced, if we use the assumption proposed in [27], we will ignore the importance of the choice of the reference image. That's because this assumption says that the predictive confidence scores are only decided by the facial image with the lower iconicity score, which means that as long as the reference facial image is not worse than the test image, the quality of the reference facial image does not influence the predictive confidence. Such a conclusion does not apply to the situation that almost all the image pairs are unbalanced. Therefore, for the unbalabced image pairs, our model based on the assumption that the final predictive confidence will be decided by both of the images in the pair makes more sense.

In order to use our iconicity score to verification task, we use our iconicity score as a filter before the input of verification system. The structure of application is showed in figure 3-6.

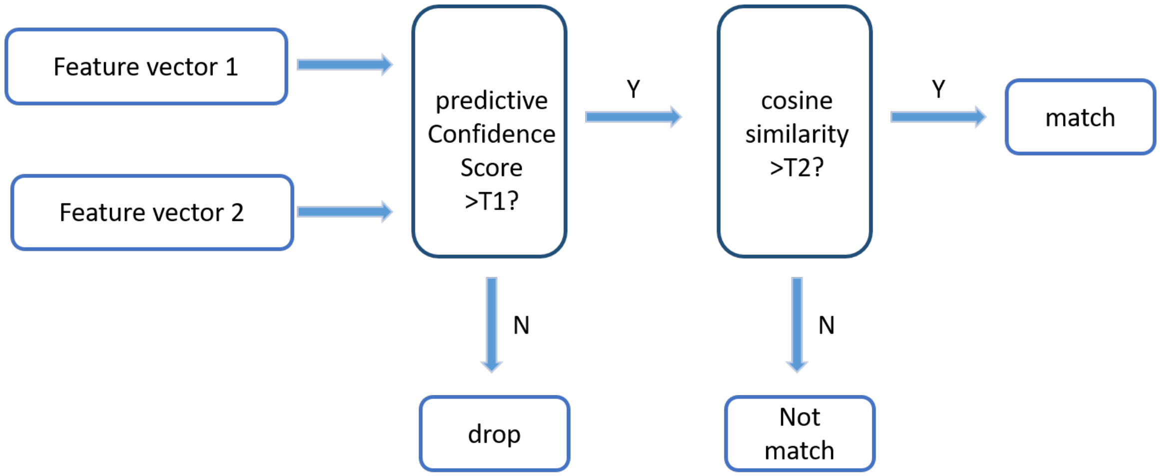


Figure 3-6. Structure of verification system

Comparison with PCNet

The goal of this metric is to perform image-to-image verification. There are 2k facial images in the test set, and we use the face detection score to divide the test set into the high-quality set and the low-quality set. While performing the verification task, the image pairs consist of one image from the high-quality set and another one from the low-quality set. We generate about 1000 image pairs in total and half of these pairs are genuine pairs (from the same individual), the other half are imposter pairs (from different individuals).

In this experiment, we rank these pairwise predictive confidence scores of image pairs in descending order and reject the bottom k percent pairs (k = 10,20,30). Then we build the ROC curves of the verification task based on the test set with the Retained proportion of 100,90,80,70 percent. The standard for image pairs to be retained is the multiple of the iconicity scores for our model and the minimum of the iconicity scores for the PCNet. The final result is shown in figure 3-7.

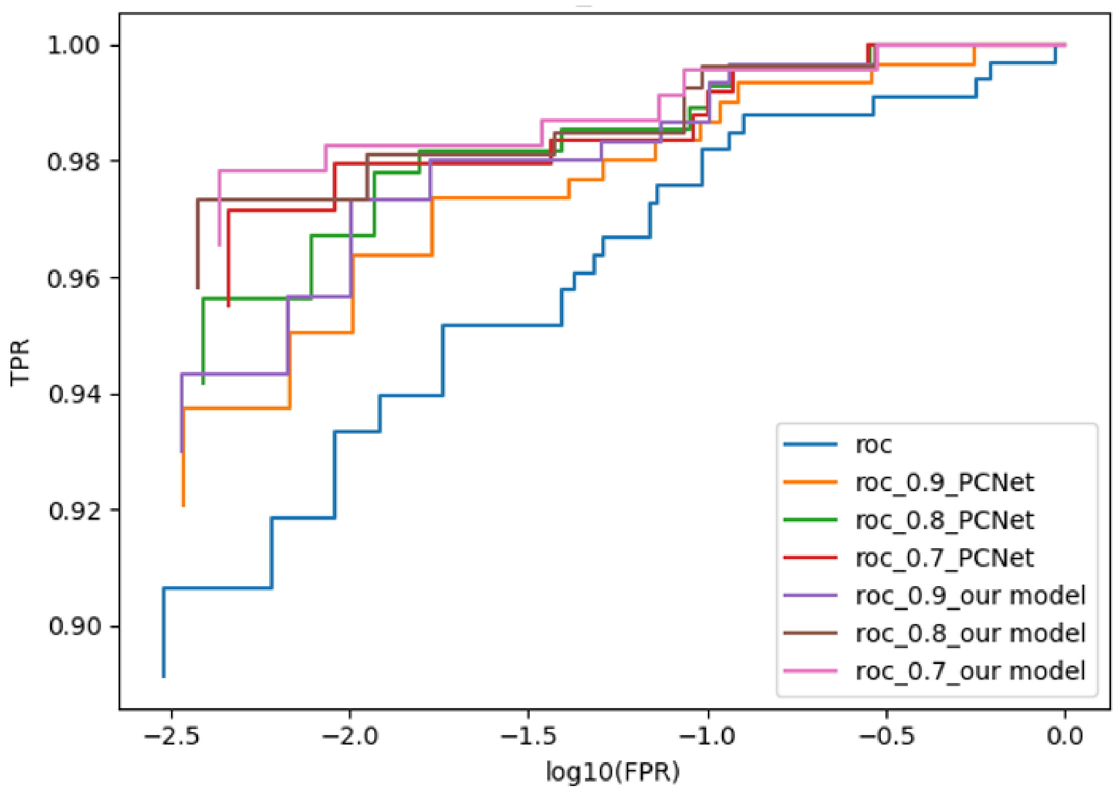


Figure 3-7. ROC curves of different model tested on image pairs with top 90,80,70 percent iconicity score

The result curves show us the performance of the verification system on the test set filtered based on the score information obtained from our Siamese MLP network and the PCNet. According to the result, we can see that with the same rejection portion, the verification task with rejection based on Siamese MLP network surpasses that based on PCNet, because the TPR of the curve with Siamese MLP network is higher than the curve with PCNet in the same FPR especially in the low-FPR-area, and such a fact happens in every rejection portion curves.

Other than rejecting the same portion of test image pairs, we also try to set a fixed threshold (here we use 0.5 to be the threshold) to reject the image pairs whose predictive confidence score is lower than the threshold and then build the ROC curve.

With the threshold of 0.5, there are 943 pairs left based on the PCNet model, and there are 927 pairs left based on our model. The result is shown in figure 3-8.

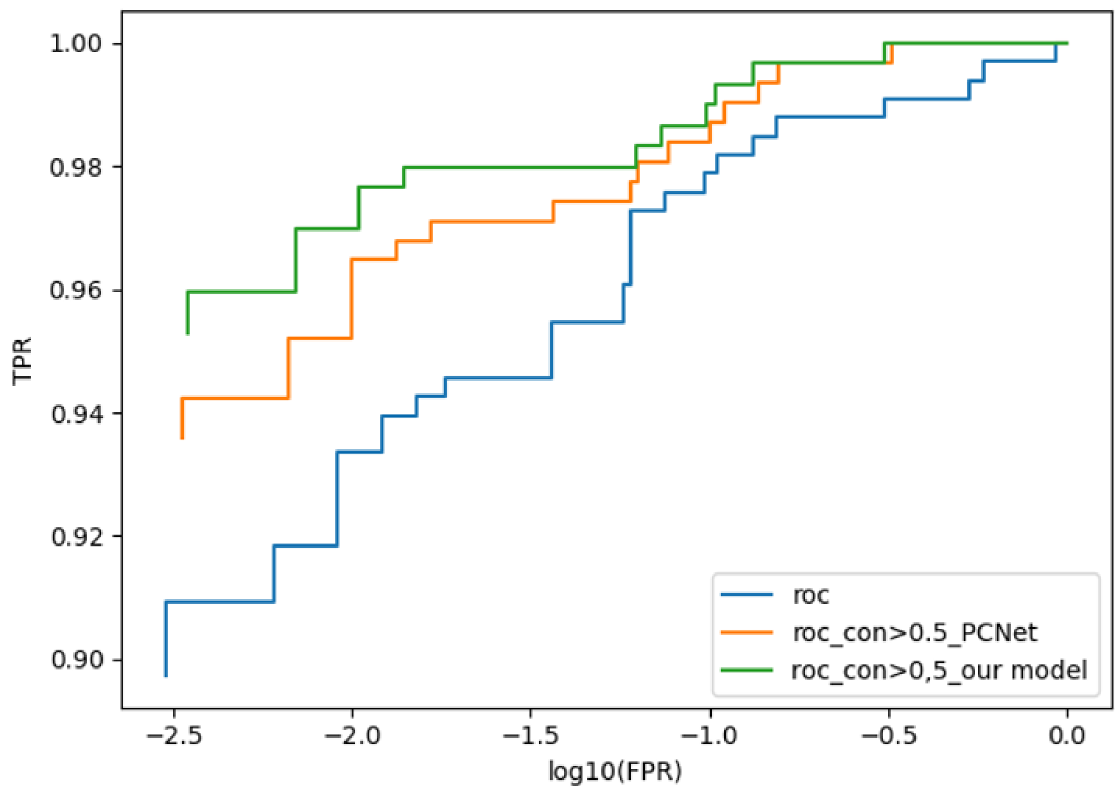


Figure 3-8. ROC curves of different models tested with image pairs whose confidence score are higher 0.5

According to the result of this experiment, we can see that by setting the same threshold on predictive confidence scores for image pairs, our Siamese MLP network performs better than the PCNet because of the higher TAR in the same FAR. As a result, we can conclude that our Siamese MLP network can help predict the failure cases better and more accurately than PCNet for verification systems. Therefore, we can imply that for the unbalanced image pairs, our Siamese MLP network can help increase the performance of face verification better than the PCNet.

Conclusions and future work

To summarize, in this project, we generate the training image pair set automatically and propose a novel training pipeline to evaluate the iconicity score of facial images. Our training process only requires the image datasets with identity information that are already available for face recognition training tasks.

3.3 Conclusions

In conclusion, in this project, we propose a method to train a network to predict the iconicity score of facial images from identity features. Our data-driven model does not need any ground truth about the quality information of the image, which can help save a lot of time and cost, and we will never be limited by the lack of the dataset with quality information. The objective of computing iconicity is to help improve the performance of the face verification system by filtering out the ineffective images in advance, such as the blurred, occluded, or even nonface images. We propose several metrics for evaluating our iconicity scores. The first one is by directly seeing the iconic and non-iconic images classified by the iconicity score obtained from our Siamese MLP network, and the result shows that the result is consistent with human intuition. The second way is to explore the relationship between the iconicity score and other visible factors such as head poses and key points. The result shows that the iconicity score correlates well with different factors which might influence the identifiability of facial images. The third one is also the original goal of our network. We check the capability of iconicity score to help increase the performance of the face verification system. Based on these metrics, we can imagine several potential applications for iconicity scores. The iconicity scores can be used to rank the quality of a set of images and help choose the best one to use. Besides, it can be used as a filter before the input of the face verification system. The filter can help filter out non-iconic images to improve the accuracy and performance of the verification system.

3.4 Future works

There are a lot of future directions in the iconicity estimation area to be explored. For example, if we obtain videos that look around ones' face/head, and we can extract the 3D information or build the depth images from these videos, it will be very useful to evaluate the iconicity score of this kind of video. Such a model may be used in some more accurate tasks and situations like 3D modeling. Besides, we can also try to evaluate the iconicity score of the whole body instead of the face. Our people recognize other ones not only by their face but also by their body features (especially in this specific period when a lot of people wearing masks all the time), therefore, we can also try to evaluate if the whole body image is informative to be recognized by machines. Moreover, we can even try to estimate the iconicity score of other dynamic characteristics of people, such as the gait.

References

1. *2017 Information technology - Vocabulary - Part 37: Biometrics* Standard (International Organization for Standardization, ISO/IEC JTC1 SC37 Biometrics, 2017).
2. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **60,** 84–90 (May 2017).
3. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* **abs/1409.1556** (2015).
4. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning for Image Recognition* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016).
5. Parkhi, O. M., Vedaldi, A. & Zisserman, A. *Deep Face Recognition* in *Proceedings of the British Machine Vision Conference (BMVC)* (eds Xianghua Xie, M. W. J. & Tam, G. K. L.) (BMVA Press, Sept. 2015), 41.1–41.12.
6. Bansal, A., Nanduri, A., Castillo, C. D., Ranjan, R. & Chellappa, R. UMDFaces: An Annotated Face Dataset for Training Deep Networks. *CoRR* **abs/1611.01484.** arXiv: 1611.01484 (2016).
7. Maze, B. *et al. IARPA Janus Benchmark - C: Face Dataset and Protocol* in *2018 International Conference on Biometrics (ICB)* (2018), 158–165.
8. Ranjan, R. *et al.* A Fast and Accurate System for Face Detection, Identification, and Verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science* **1,** 82–96 (2019).
9. Chen, J., Deng, Y., Bai, G. & Su, G. Face Image Quality Assessment Based on Learning to Rank. *IEEE Signal Processing Letters* **22,** 90–94 (2015).
10. Zhang, L., Zhang, L. & Li, L. *Illumination Quality Assessment for Face Images: A Benchmark and a Convolutional Neural Networks Based Model* in *Neural Information Processing* (eds Liu, D., Xie, S., Li, Y., Zhao, D. & El-Alfy, E.-S. M.) (Springer International Publishing, Cham, 2017), 583–593.
11. Yang, F. *et al.* in, 655–667 (Nov. 2019).
12. Best-Rowden, L. & Jain, A. K. Automatic Face Image Quality Prediction. *CoRR* **abs/1706.09887.** arXiv: 1706.09887 (2017).
13. Best-Rowden, L. & Jain, A. K. Learning Face Image Quality From Human Assessments. *IEEE Transactions on Information Forensics and Security* **13,** 3064–3077 (2018).

14. Dhar, P., Castillo, C. & Chellappa, R. *On Measuring the Iconicity of a Face* in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2019), 2137–2145.

15. Natu, V. S. & O'Toole, A. J. Spatiotemporal changes in neural response patterns to faces varying in visual familiarity. *NeuroImage* **108,** 151–159 (2015).

16. Vaidehi Natu, A. O. *Spatial and temporal characteristics of the neural representation of face familiarity* Vision Sciences Society Annual Meeting. 2012.

17. Singh, M., Singh, R. & Ross, A. A comprehensive overview of biometric fusion. *Information Fusion* **52,** 187–205 (2019).

18. Fierrez, J., Morales, A., Vera-Rodriguez, R. & Camacho, D. Multiple classifiers in biometrics. Part 2: Trends and challenges. *Information Fusion* **44,** 103–112 (2018).

19. *Information technology - Biometric data interchange formats - Part 5: Face image data* Standard (International Organization for Standardization, ISO/IEC JTC1 SC37 Biometrics, 2011).

20. *Machine Readable Passports – Part 9 – Deployment of Biometric Identification and Electronic Storage of Data in eMRTDs* Standard (International Civil Aviation Organization, ISO/IEC JTC1 SC17 WG3, 2018).

21. *Portrait Quality - Reference Facial Images for MRTD* Standard (International Civil Aviation Organization, https://www.icao.int/publications/Documents/9303$_p$9$_c$*ons*$_e$*n.pdf*, 2015).

22. *2010 Information technology - Biometric sample quality - Part 5: Face image data* Standard (International Organization for Standardization, ISO/IEC JTC1 SC37 Biometrics, 2010).

23. Abaza, A., Harrison, M. A. & Bourlai, T. *Quality metrics for practical face recognition* in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)* (2012), 3103–3107.

24. Mittal, A., Moorthy, A. K. & Bovik, A. C. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing* **21,** 4695–4708 (2012).

25. Parde, C. J. *et al. Face and Image Representation in Deep CNN Features* in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)* (2017), 673–680.

26. Yang, J. *et al. Neural Aggregation Network for Video Face Recognition* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017).

27. Xie, W., Byrne, J. & Zisserman, A. *Inducing Predictive Uncertainty Estimation for Face Recognition* 2020. arXiv: 2009.00603 [cs.CV].

28. Maze, B. *et al.* IARPA Janus Benchmark - C: Face Dataset and Protocol. *2018 International Conference on Biometrics (ICB),* 158–165 (2018).

29. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *CoRR* **abs/1512.03385.** arXiv: 1512.03385 (2015).

30. Ranjan, R., Sankaranarayanan, S., Castillo, C. D. & Chellappa, R. An All-In-One Convolutional Neural Network for Face Analysis. *CoRR* **abs/1611.00851.** arXiv: 1611.00851 (2016).

31. Hernandez-Ortega, J., Galbally, J., Fierrez, J., Haraksim, R. & Beslay, L. *FaceQnet: Quality Assessment for Face Recognition based on Deep Learning* in *2019 International Conference on Biometrics (ICB)* (2019), 1–8.

32. Xie, W. & Zisserman, A. *Multicolumn Networks for Face Recognition* 2018. arXiv: 1807.09192 [cs.CV].